CAPITOLO 6

SISTEMI LINEARI

In questo capitolo ci occuperemo delle risoluzione numerica dei sistemi lineari del tipo

$$Ax=b$$

con $A \in \mathcal{L}(C^n, C^n)$, $x,b \in C^n$, ed $|A| \neq 0$.

I sistemi lineari rappresentano dei modelli per una vasta gamma di problemi attinenti alle scienze applicate ed all'ingegneria quali il calcolo delle reti elettriche, la dinamica discreta di popolazioni, il calcolo delle strutture statiche ecc. Spesso i sistemi derivanti da tali problemi sono di dimensioni moderate, e non presentano difficoltà di risoluzione dal punto di vista numerico anche se le matrici che intervengono sono generalmente matrici piene e non strutturate. Per tali sistemi saranno generalmente sufficienti i metodi diretti che verranno esposti nel paragrafo 2.

Le più grosse difficoltà sorgono invece nella risoluzione dei sistemi lineari che si ottengono dalla discretizzazione di certi sistemi di equazioni differenziali, ordinarie o a derivate parziali. Se la discretizzazione è abbastanza fine, allora essa può dar luogo ad un sistema lineare di dimensione molto grande. D'altra parte la particolare natura delle equazioni differenziali insieme a particolari scelte delle discretizzazioni, danno luogo generalmente a sistemi lineari di forma particolare che consentono l'uso di certe tecniche iterative che sono presentate e studiate nel paragrafo 1 di questo capitolo.

A tale scopo onsideriamo due tipici problemi differenziali del secondo ordine (uno alle derivate ordinarie ed uno alle derivate parziali) e una loro possibile discretizzazione.

Osserviamo preliminarmente che la derivata seconda u"(t) di una funzione può essere approssimata con la seguente *differenza centrale*

,,

$$u''(t) \approx \frac{u(t-h) - 2u(t) + u(t+h)}{h^2}$$

Sommando gli sviluppi in serie di Taylor fino al quarto ordine di u(t-h) e u(t+h) si ottiene infatti:

$$u''(t) = \frac{u(t-h) - 2u(t) + u(t+h)}{h^2} + O(h^2)$$

Problema dei due punti.

Data una funzione $g(t) \geq 0$ in $[0,1]$, si consideri , per una assegnata $f(t)$, l'equazione differenziale

$$u''(t) - g(t)u(t) = f(t) \qquad t \in [0,1]$$

con le condizioni agli estremi:

$$u(0)= a, \qquad u(1)= b.$$

Si può dimostrare che tale problema ammette una ed una sola soluzione per ogni termine noto $f(t)$.
Consideriamo ora una discretizzazione dell'intervallo $[0,1]$ di passo $h=1/N$, con N intero,

$$t_i=ih \quad i=0,1,... N$$

e indichiamo con u_i le approssimazioni di $u(t_i)$, e con g_i ed f_i i valori $g(t_i)$ e $f(t_i)$ rispettivamente. Approssimando $u''(t_i)$, per $i=1,...N-1$, con la differenza centrale $\dfrac{u(t-h) - 2u(t) + u(t+h)}{h^2}$ si ottiene,cambiando segno all'equazione e tenendo conto che $u_0=a$ ed $u_N=b$, il seguente sistema lineare:

$$(2 + h^2g_1)u_1 - u_2 = -h^2f_1 + a$$
$$.....$$
$$-u_{i-1} +(2 + h^2g_i)u_i - u_{i+1} = -h^2f_i$$
$$.....$$
$$-u_{N-2} +(2 + h^2g_{N-1})u_{N-1}= -h^2f_{N-1}+ b$$

la cui matrice :

$$A = \begin{Vmatrix} 2 + h^2 g_1 & -1 & & & & \\ -1 & 2 + h^2 g_2 & -1 & & & \\ & . & . & . & & \\ & & -1 & 2 + h^2 g_i & -1 & \\ & & & . & . & . \\ & & & & -1 & 2 + h^2 g_{N-2} & -1 \\ & & & & & -1 & 2 + h^2 g_{N-1} \end{Vmatrix}$$

risulta simmetrica e tridiagonale.
Dimostriamo ora che la matrice A risulta anche definita positiva. A tale scopo osserviamo che

$$x^H A x = x^H T x + x^H G x$$

dove :

$$T = \begin{Vmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & . & . & . & & \\ & & . & . & . & \\ & & & . & . & . \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{Vmatrix} \qquad G=\mathrm{diag}(h^2 g_1,\ldots h^2 g_i \ldots, h^2 g_{N-1})$$

Poichè la matrice G è evidentemente semidefinita positiva, a seguito della condizione $g(t) \geq 0$, A risulterà definita positiva se tale è T.

La dimostrazione che T è definita positiva si può ottenere in maniera diretta sviluppando la forma quadratica $x^T T x$. Calcolando prima il vettore Tx e successivamente il prodotto scalare $x^T(Tx)$, si ottiene facilmente:

$$x^T T x = 2x_1^2 - 2x_1 x_2 + 2x_2^2 - 2x_2 x_3 + 2x_3^2 - 2x_3 x_4 + \ldots - 2x_{n-1} x_n + 2x_n^2 =$$

$$x_1^2 + (x_1^2 - 2x_1 x_2 + x_2^2) + (x_2^2 - 2x_2 x_3 + x_3^2) + \ldots + (x_{n-1}^2 - 2x_{n-1} x_n + x_n^2) + x_n^2 =$$

$$x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + \ldots + (x_{n-1} - x_n)^2 + x_n^2$$

Tale numero è evidentemente non negativo, ed è positivo per ogni vettore non identicamente nullo.
La matrice T è dunque definita positiva e ciò assicura, come abbiamo osservato prima, che anche A è definita positiva.

Problema di Dirichelet.

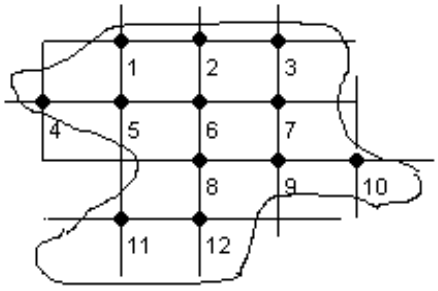
Sia u(x,y) una funzione definita in un dominio D del piano R^2, il cui bordo sarà indicato con Γ_D , che soddisfa l'equazione a derivate parziali

$$\frac{\partial^2 u}{\partial^2 x} + \frac{\partial^2 u}{\partial^2 y} = f(x,y) \qquad x,y \in D$$

con la *condizione al contorno*,

$$u(x,y)=0 \qquad x,y \in \Gamma_D.$$

Effettuiamo una reticolazione del dominio D con rette parallele agli assi coordinati, distanti tra loro una quantità positiva h, ed enumeriamo con un indice progressivo i nodi P_i che risultano *interni* al dominio.



In riferimento alla discretizzazione in figura, ci sono 12 punti, P_i i=1,…,12, interni al dominio per i quali si cercano i valori $u(P_i)$.
Usando le differenze centrali per approssimare le derivate parziali seconde, si ottiene:

$$\frac{\partial^2 u(P_i)}{\partial^2 x} + \frac{\partial^2 u(P_i)}{\partial^2 y} \approx \frac{u(P_{iw}) - 2u(P_i) + u(P_{ie})}{h^2} + \frac{u(P_{in}) - 2u(P_i) + u(P_{is})}{h^2}$$

e quindi

$$\frac{\partial^2 u(P_i)}{\partial^2 x} + \frac{\partial^2 u(P_i)}{\partial^2 y} \approx \frac{u(P_{in}) + u(P_{iw}) - 4u(P_i) + u(P_{ie}) + u(P_{is})}{h^2}$$

dove i punti P_{iw}, P_{ie}, P_{in}, P_{is}, rappresentano rispettivamente il punto a sinistra (west), a destra (east), sopra (north) e sotto (south) al punto P_i.
Se uno di questi punti e' esterno al dominio o giace sul dominio, viene posto uguale a zero (come se fosse sul bordo del dominio stesso). Se invece e' interno al dominio, assume l'indice che gli compete in accordo con la geometria del dominio e con la discretizzazione adottata.

Per esempio nel punto P_1 della nostra figura si ha

$$\frac{\partial^2 u(P_1)}{\partial^2 x} + \frac{\partial^2 u(P_1)}{\partial^2 y} \approx \frac{-4u(P_1) + u(P_2) + u(P_5)}{h^2}$$

mentre nel punto P_8 si ha

$$\frac{\partial^2 u(P_8)}{\partial^2 x} + \frac{\partial^2 u(P_8)}{\partial^2 y} \approx \frac{u(P_6) - 4u(P_8) + u(P_9) + u(P_{12})}{h^2}$$

In riferimento alla figura, il sistema lineare nelle incognite $u(P_i)$ i=1,...,12 ha una matrice ***tridiagonale a blocchi*** della forma:

$$A = \left\|\begin{array}{cccccccccccc}
4 & -1 & & -1 & & & & & & & & \\
-1 & 4 & -1 & & -1 & & & & & & & \\
 & -1 & 4 & & & -1 & & & & & & \\
 & & & 4 & -1 & & & & & & & \\
-1 & & & -1 & 4 & -1 & & & & & & \\
 & -1 & & & -1 & 4 & -1 & & & & & \\
 & & -1 & & & -1 & 4 & & & & & \\
 & & & & & & & 4 & -1 & & -1 & \\
 & & & & & & -1 & -1 & 4 & -1 & & \\
 & & & & & & & & -1 & 4 & & \\
 & & & & & & & & & & 4 & -1 \\
 & & & & & & & -1 & & & -1 & 4
\end{array}\right\| \qquad (2.1)$$

E' facile rendersi conto che la matrice risulta simmetrica. Inoltre, con ragionamenti simili all'esempio precedente, si dimostra anche che è definita positiva. E' anche facile capire che se l'equazione fosse stata di dimensione 3, cioè:

$$\frac{\partial^2 u}{\partial^2 x} + \frac{\partial^2 u}{\partial^2 y} + \frac{\partial^2 u}{\partial^2 z} = f(x,y,z) \qquad x,y,z \in D$$

il sistema sarebbe risultato *pentadiagonale a blocchi* e la sua dimensione sarebbe cresciuta considerevolmente. Se il dominio D fosse, per esempio, un cubo con lo spigolo unitario, un reticolo con 10 punti interni su ogni coordinata darebbe luogo ad un sistema di dimensione 1000.

1.METODI ITERATIVI.

I metodi iterativi per la risoluzione dei sistemi lineari si fondano sulla trasformazione dell'equazione $Ax=b$ in un problema equivalente di punto fisso. Ciò si ottiene attraverso uno *spezzamento* (*splitting*) della matrice A in

$A = N - P$, con N non singolare e $P=N-A$,

che dà luogo a

$$Nx = Px + b$$
$$x= N^{-1}Px + N^{-1}b$$

Il problema $Ax=b$ è cosi trasformato nel problema di punto fisso

$$x = Mx + a \qquad (2.2)$$

dove $a=N^{-1}b$ ed $M = N^{-1}P = I-N^{-1}A$. La matrice M è detta *matrice di iterazione* ed è definita univocamente dallo splitting.

Il problema di punto fisso viene quindi affrontato con l'iterazione

$$x^{k+1}=Mx^k +a, \qquad (2.3)$$

o, più precisamente, con l'iterazione

$$Nx^{k+1} = Px^k + b$$

a partire da un vettore iniziale x^0 assegnato. Affinchè questo approccio possa essere vantaggioso occorre che il sistema

$$Ny = Px^k + b$$

sia risolvibile per y in maniera "diretta", cioè con un costo trascurabile rispetto al costo richiesto per la risoluzione del problema originale Ax=b.

Sottraendo (2.2) da (2.3) si ottiene, per gli errori $e^k := x^k - x$,

$$e^{k+1} = Me^k = M^2e^{k-1} =... = M^{k+1}e^0$$

Se la matrice M è convergente allora l'errore tende a zero, qualunque sia il vettore iniziale x^0, mentre può accadere che, per un certo x^0, l'errore tenda a zero senza che la matrice sia infinitesima. Ciò accade quando M^k tende ad una matrice il cui nucleo includa e^0. Se invece si vuole che l'errore tenda a zero per ogni vettore iniziale, allora M deve essere convergente. Abbiamo dunque dimostrato il seguente teorema.

TEOREMA 2.1. *Condizione necessaria e sufficiente affinchè l'iterazione* (2.3) *converga per ogni vettore iniziale* x^0 *è che la matrice* M *sia infinitesima o, equivalentemente, che* $\rho(M)<1$.

Dunque nel proporre un metodo iterativo, lo splitting deve essere scelto con i seguenti criteri:

1) N deve essere non singolare.
2) N deve essere invertibile a costo trascurabile.
3) La matrice di iterazione M che ne deriva deve essere convergente.
4) Il raggio spettrale di M deve essere più piccolo possibile.

Quest'ultima affermazione discende dalla seguente analisi asintotica dell'errore.

Analisi asintotica dell'errore.

Nel valutare la bontà di un metodo iterativo attraverso l'analisi della successione e^k degli errori , si deve osservare che questa dipende dal punto iniziale x^0 col quale si innesca l'iterazione e dalla particolare norma con la quale viene valutata l'ampiezza degli

errori. Una valutazione corretta della velocità di convergenza del metodo deve prescindere da entrambi questi fattori. Partendo quindi dalla relazione ricorsiva sugli errori

$$e^n = Me^{n-1} = M^n e^0$$

si ottiene, per una norma arbitraria,:

$$||e^n|| \leq ||M^n|| ||e^0||$$

e quindi,

$$\frac{||e^n||}{||e^0||} \leq ||M^n||$$

Dunque in n iterazioni il fattore di smorzamento dell'errore $\frac{||e^n||}{||e^0||}$ è maggiorato da $||M^n||$, e quindi la quantità $\sqrt[n]{\frac{||e^n||}{||e^0||}}$, che rappresenta, dopo n passi, il fattore medio di smorzamento ad ogni passo, è a sua volta maggiorato da

$$\sqrt[n]{\frac{||e^n||}{||e^0||}} \leq \sqrt[n]{||M^n||} \ .$$

Il termine a destra della disuguaglianza ammette limite, per $n \to \infty$, pari a $\rho(M)$. Di conseguenza il termine a sinistra, che in generale non ammette limite, possiede certamente il massimo limite e per esso si avrà:

$$\max \lim_{n\to\infty} \sqrt[n]{\frac{||e^n||}{||e^0||}} \leq \rho(M).$$

Osserviamo infine che scegliendo x^0 in modo tale che e^0 sia autosoluzione di M associata all'autovalore μ di modulo massimo, si ha

$$e_n = M^n e_0 = \mu^n e_0$$

Da ciò si ricava

$$||e^n|| = \rho^n(M) ||e^0||$$

e quindi

$$\sqrt[n]{\frac{\|e^n\|}{\|e^0\|}}=\rho(M).$$

Di conseguenza rispetto a tutti i possibili vettori iniziali ed a tutte le norme si ha:

$$\max_{e^0}\left(\max \lim_{n\to\infty} \sqrt[n]{\frac{\|e^n\|}{\|e^0\|}}\right)= \rho(M)$$

Il termine a sinistra si chiama **fattore asintotico di convergenza** e rappresenta il peggiore fattore medio di cui viene smorzato asintoticamente l'errore iniziale, indipendentemente dal vettore iniziale e dalla norma con cui si misura l'errore.

Il fatto che il fattore asintotico di convergenza coincide col raggio spettrale della matrice di iterazione del metodo, giustifica l'asserzione che un metodo iterativo è, *in generale,* tanto più veloce quanto più piccolo è il raggio spettrale. Ciò non esclude che per certi vettori iniziali un metodo risulti più veloce di un altro di raggio spettrale inferiore.

Per dare una stima a priori del numero di iterazioni necessarie per avere un errore al di sotto di una assegnata tolleranza, è utile il concetto di **ordine asintotico di convergenza**. Esso rappresenta il numero di iterazioni necessarie per smorzare asintoticamente l'errore di un fattore 10^{-1} e verrà indicato con R.

Sulla base dei risultati precedenti si può dire che asintoticamente, cioè dopo un numero sufficientemente grande di iterazioni, sarà sufficiente che R soddisfi la disuguaglianza:

$\rho^R \le 10^{-1}$ e quindi , poiche $\rho<1$, $R\ge -1/\log_{10}\rho$

Come prima, questa è una stima pessimistica inquanto il fattore asintotico di smorzamento per l'iterazione che si sta calcolando potrebbe essere, in realtà, più piccolo. Si osservi che se un altro metodo ha il raggio spettrale che è il quadrato di ρ, il suo ordine asintotico di convergenza è la metà. Si noti infine che disponendo di una maggiorazione del raggio spettrale, purchè inferiore ad 1, si dispone di una maggiorazione di R.

Metodo di Jacobi (metodo delle sostituzioni simultanee).

Il metodo di Jacobi consiste nel decomporre la matrice A in

$N=D:=\mathrm{diag}(a_{11},....,a_{nn})$ e $P=N-A$.

Poichè N deve essere invertibile, si deve premoltiplicare A per una matrice di permutazione in modo che la matrice permutata abbia sulla diagonale elementi tutti non nulli. Si vede che ciò è sempre possibile a condizione che A stessa sia non singolare.

Risulta $N^{-1} = \mathrm{diag}(\frac{1}{a_{11}},\ldots,\frac{1}{a_{nn}})$ e per la matrice di iterazione $M=(m_{ij})$ si ha:

$$m_{ij} = -\frac{a_{ij}}{a_{ii}}, \quad \text{per } i \neq j, \quad \text{ed } m_{ii}=0.$$

L'iterazione di Jacobi si scrive dunque:

$$x_i^{k+1} = -\sum_{j\neq i} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}} \qquad i=1,\ldots,n$$

Una matrice A si dice **a predominanza diagonale stretta**, se per ogni riga si ha:

$$\sum_{j\neq i} |a_{ij}| < |a_{ii}| \qquad (i=1,\ldots,n) .$$

Poichè la predominanza diagonale stretta implica $\sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1$ per ogni i, allora si dimostra immediatamente il seguente teorema

Teorema 2.2. *Se la matrice A è a predominanza diagonale stretta allora il metodo di Jacobi è convergente*.

Dim: $\| M \|_\infty = \max_i \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$

Un'altra condizione sufficiente per la convergenza è data da $\| M \|_1 < 1$,
cioè da

$$\max_j \sum_{i\neq j} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Si badi bene che la precedente relazione non significa la **predominanza diagonale per colonne**, che si esprime invece con $\sum_{i\neq j} |a_{ij}| < |a_{jj}| \qquad (j=1,\ldots,n)$, ma una relazione più

complessa che può comunque essere utile in qualche caso. D'altra parte la predominanza diagonale per colonne assicura anch'essa la convergenza del matodo di Jacobi.

Teorema 2.3. *Se la matrice A è a predominanza diagonale stretta per colonne, il metodo di Jacobi converge*.

Dim. Data la matrice B, indichiamo con B^{-T} la trasposta dell'inversa di B. La predominanza diagonale per colonne di A implica la predominanza diagonale per righe di A^T, la cui matrice di iterazione di Jacobi è

$$M'=D^{-T}P^T=D^{-1}P^T.$$

Per tale matrice si ha, in base al teorema precedente, $\rho(D^{-1}P^T)<1$ e quindi anche $\rho((D^{-1}P^T)^T)=\rho(PD^{-1})<1$. Poichè PD^{-1} è simile a $D^{-1}P$ anche $\rho(D^{-1}P)<1$.

Metodo di Gauss-Seidel (metodo delle sostituzioni successive).

Un'altra decomposizione di A in N-P, con N facilmente invertibile, è data da N=L+D e P=-U dove D è la diagonale ed L e U sono la parte triangolare inferiore e superiore di A. L'iterazione $Nx^{k+1} = Px^k + b$ assume la forma

$$\begin{aligned}
&a_{11}\, x_1^{k+1} && = -a_{12}\, x_2^{k} - a_{13}\, x_3^{k} - \ldots\ldots - a_{1n}\, x_n^{k} + b_1 \\
&a_{21}\, x_1^{k+1} + a_{22}\, x_2^{k+1} && = \quad - a_{23}\, x_3^{k} - \ldots\ldots - a_{2n}\, x_n^{k} + b_2 \qquad \ldots\ldots \\
&\ldots\ldots \\
&a_{n1}\, x_1^{k+1} + a_{n2}\, x_2^{k+1} + \ldots\ldots + a_{nn}\, x_n^{k+1} && = \qquad\qquad + b_n
\end{aligned}$$

Tale sistema si risolve facilmente ed ogni componente x_i^{k+1} è data da:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{k} + \frac{b_i}{a_{ii}} \qquad i=1,\ldots,n$$

Come si può osservare, nel metodo delle sostituzione successive il valore aggiornato di ogni singola componente x_j^{k+1} viene utilizzato immediatamente per il calcolo delle componenti successive relative alla stessa iterazione, mentre nel metodo delle sostituzioni simultanee i valori aggiornati di tutte le componenti vengono sostituiti simultaneamente ai valori x_j^k alla fine dell'iterazione. A differenza del metodo di Jacobi, ora non è immediato

valutare la matrice di iterazione $M = N^{-1}P = (L+D)^{-1}(-U)$ e quindi una sua norma, di conseguenza la nostra analisi della convergenza sarà effettuata analizzando l'errore componente per componente. A tale scopo, osservato che la soluzione del sistema $Nx=Px + b$ si può scrivere

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}} \qquad i=1,...,n$$

si ottiene, per gli errori $e_j^k := x_j^k - x_j$

$$e_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^k$$

$$|e_i^{k+1}| \le \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{k+1}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^k| \qquad i=1,...,n$$

$$\left|e_i^{k+1}\right| \le \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \|e^{k+1}\|_\infty + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \|e^k\|_\infty \qquad i=1,...,n.$$

Detto M l'indice per il quale il secondo termine della disuguaglianza è massimo, e chiamato per brevità $R_M = \sum_{j=1}^{M-1} \left|\frac{a_{Mj}}{a_{MM}}\right|$ ed $S_M = \sum_{j=M+1}^{n} \left|\frac{a_{Mj}}{a_{MM}}\right|$, si ha

$$\left|e_i^{k+1}\right| \le R_M \|e^{k+1}\|_\infty + S_M \|e^k\|_\infty \quad \text{per ogni } i$$

e quindi

$$\|e^{k+1}\|_\infty \le R_M \|e^{k+1}\|_\infty + S_M \|e^k\|_\infty$$

$$(1-R_M)\|e^{k+1}\|_\infty \le S_M \|e^k\|_\infty$$

Se supponiamo A a predominanza diagonale stretta, allora $R_M+S_M <1$, da cui segue che $R_M <1$ ed $S_M <1$ e quindi

$$\|e^{k+1}\|_\infty \le \frac{S_M}{1-R_M} \|e^k\|_\infty \le \left(\frac{S_M}{1-R_M}\right)^{k+1} \|e^0\|_\infty$$

dove, per la predominanza diagonale, è ancora $\frac{S_M}{1-R_M}<1$. Abbiamo dimostrato dunque il seguente teorema.

Teorema 2.4. *Se la matrice A è a predominanza diagonale stretta allora il metodo di Gauss-Seidel è convergente*.

Sebbene la predominanza diagonale sia sufficiente per la convergenza di entrambi i metodi, esistono matrici A per le quali un metodo converge e l'altro no. In certi casi si può tuttavia fare un confronto tra il metodo di Jacobi e quello di Gauss-Seidel. Vale infatti il seguente teorema che viene enunciato senza dimostrazione.

Teorema 2.5. *Se la matrice* A *è tridiagonale , allora per le matrici di iterazione di Jacobi e di Gauss-Seidel vale la seguente relazione* $\rho(M_{GS})=\rho^2(M_J)$.

Il teorema rivela che i due metodi convergono o divergono entrambi e, quando convergono, il metodo di Gauss-Seidel è più veloce.

Nel caso di matrici A hermitiane e definite positive, si può dare il seguente criterio per la convergenza di un metodo iterativo generato da uno splitting.

Lemma 2.6: *Sia* A *hermitiana e definita positiva e sia* N *una matrice non singolare tale che* $Q:=N + N^H - A$ *sia ancora definita positiva. Allora la matrice di iterazione* $M=I-N^{-1}A$ *è convergente*.

Dim Sia λ un autovalore di M ed u una corrispondente autosoluzione. Allora:

$Mu=\lambda u$
$NMu=\lambda Nu$
$N(I-N^{-1}A)u=\lambda Nu$
$(N-A)u=\lambda Nu$
$(1-\lambda)Nu=Au$

Poichè A è non singolare, $Au\neq 0$ e quindi $(1-\lambda)\neq 0$, per cui:

$$Nu=\frac{Au}{1-\lambda}$$
$$u^HNu = \frac{u^HAu}{1-\lambda}$$

e, passando ai coniugati,

$$u^H N^H u = \frac{u^H A u}{1-\bar{\lambda}} .$$

Sommando infine le ultime due relazioni:

$$u^H (N^H + N) u = u^H A u \left(2\, \mathcal{R}e\left(\frac{1}{1-\lambda}\right)\right)$$

$$\left(2\, \mathcal{R}e\left(\frac{1}{1-\lambda}\right)\right) = \frac{u^H (N^H + N) u}{u^H A u} = \frac{u^H (Q + A) u}{u^H A u} = \frac{u^H Q u}{A u} + 1 > 1.$$

Sia $\lambda = \alpha + i\beta$, allora

$$\frac{1}{1-\lambda} = \frac{1-\alpha+i\beta}{(1-\alpha)^2+\beta^2}$$

da cui:

$$2\, \mathcal{R}e\left(\frac{1}{1-\lambda}\right) = \frac{2(1-\alpha)}{(1-\alpha)^2+\beta^2}$$

e la condizione $2\, \mathcal{R}e\left(\frac{1}{1-\lambda}\right) > 1$ diventa $\alpha^2+\beta^2<1$, cioè $|\lambda|<1$.

Con questo criterio, che non appare utile per la convergenza del metodo di Jacobi, si puo dare il seguente teorema per la convergenza del metodo di Gauss-Seidel.

Teorema 2.7. *Se la matrice A è hermitiana e definita positiva, il metodo di Gauss-Seidel è convergente*.

Dim. Per il metodo di Gauss-Seidel si ha $N=L+D$ e, poichè A è hermitiana, $A=L+D+L^H$. Quindi $Q=N + N^H - A = L+D+L^H+D-(L+D+L^H) = D$ che risulta definita positiva avendo tutti gli elementi positivi.

Osservazione. Il teorema precedente, pur avendo delle ipotesi abbastanza restrittive, può essere molto utile in casi più generali appena si osservi che il sistema $Ax=b$,con A non singolare, è equivalente al sistema $A^HAx=A^Hb$ la cui matrice A^HA è hermitiana e definita positiva. Nell'adottare questa strategia si deve, però, tener conto del rischio dovuto al fatto

che il condizionamento della matrice A^HA è peggiore di quello di A (si veda il capitolo sul condizionamento).

Sulla base dei teoremi appena esposti, possiamo dire che il metodo di Gauss-Seidel è convergente per entrambi i sistemi generati dai problemi presentati all'inizio del capitolo, le cui matrici A sono simmetriche e definite positive. In particolare per il problema dei due punti, poichè la matrice A è anche tridiagonale, il metodo di Jacobi è ancora convergente ma e' piu' lento. Piu' precisamente l'ordine di convergenza e' doppio .

Si dimostra che il raggio spettrale della matrice di iterazione di Jacobi relativa al sistema tridiagonale $Tx=b$ è $\rho(M_J)=\cos(\pi/(N+1))$ dove N è la dimensione di T. Osserviamo che per $N=10$ si ha $\rho(M_J)\approx 0.9595$ e quindi $\rho(M_{GS})\approx 0.9206$ che rivela una velocità di convergenza molto lenta anche per il metodo di Gauss-Seidel. E' necessario quindi cercare degli ulteriori metodi, o delle modifiche ai metodi visti, che siano più veloci.

Il metodo SOR (successive over-relaxation method).

Il metodo SOR è una modifica del metodo di Gauss-Seidel e consiste nell'assumere come valore aggiornato della componente i-esima non il valore fornito dal metodo di Gauss-Seidel, che ora indichiamo con:

$$\bar{x}_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{k} + \frac{b_i}{a_{ii}}$$

bensì il valore

$$x_i^{k+1} = x_i^{k} + \omega\,(\bar{x}_i^{k+1} - x_i^{k}).$$

In altre parole si incrementa o si riduce il salto che il metodo di Gauss-Seidel provocherebbe alla componente i-esima con un opportuno **parametro di rilassamento** ω. Siccome per $\omega=1$ si ritrova il metodo di Gauss-Seidel, è ragionevole sperare che per valori diversi da 1 si abbia un metodo più veloce.

Fatte le opportune sostituzioni si trova

$$x_i^{k+1} = x_i^{k} - \omega\, x_i^{k} + \omega \left(-\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{k} + \frac{b_i}{a_{ii}} \right)$$

da cui si ricava:

$$a_{ii}x_i^{k+1} + \omega \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} = a_{ii}(1-\omega)x_i^k - \omega \sum_{j=i+1}^{n} a_{ij}x_j^k + \omega b_i$$

che rappresenta la riga i-esima del seguente sistema:

$$Dx^{k+1} + \omega Lx^{k+1} = D(1-\omega)x^k - \omega Ux^k + \omega b$$

$$(D + \omega L)x^{k+1} = \big((1-\omega)D - \omega U\big)x^k + \omega b.$$

Dividendo per ω , si ottiene lo splitting

$$\left(\frac{D}{\omega} + L\right)x^{k+1} = \left(\frac{1-\omega}{\omega}D - U\right)x^k + b$$

ed il metodo iterativo si esprime con

$$x^{k+1} = H_\omega x^k + \omega(D + \omega L)^{-1}b$$

dove la matrice di iterazione è:

$$H_\omega = (D + \omega L)^{-1}\big((1-\omega)D - \omega U\big).$$

Si noti che mentre nel metodo di Gauss-Seidel tutta la diagonale D appartiene alla parte N dello splitting, in SOR la parte $\frac{D}{\omega}$ sta in N mentre $\frac{1-\omega}{\omega}D$ sta in P.

Si tratta ora di vedere per quali valori del parametro di rilassamento si ha $\rho(H_\omega)<1$ ma, soprattutto, per quali valori di ω si ha un metodo più veloce di Gauss-Seidel ed in particolare quale è il valore ottimale di ω che minimizza $\rho(H_\omega)$.

Teorema 2.8 (di Kahan). *Per ogni matrice* A *tale che* $|D|\neq 0$, *si ha:*

$$\det(H_\omega) = (1-\omega)^n$$

$$\rho(H_\omega) \geq |1-\omega|.$$

Dim. Poichè il determinante di una matrice triangolare, o della sua inversa, dipende solo dagli elementi diagonali, si ha:

$$\det(H_\omega) = \det(D + \omega L)^{-1} \det\big((1-\omega)D - \omega U\big) = \det D^{-1} \det\big((1-\omega)D\big) =$$
$$= \det\big((1-\omega)I\big) = (1-\omega)^n$$

$\rho(H_\omega) \geq |1-\omega|$ discende immediatamente dal fatto che il determinante di una matrice è il prodotto dei suoi autovalori.

Corollario 2.9. *Condizione necessaria affinchè* $\rho(H_\omega)<1$ *è* $0<\omega<2$.

Dim. $\rho(H_\omega)<1 \Rightarrow |1-\omega|<1 \Rightarrow 0<\omega<2$.

Nel caso di matrici hermitiane e definite positive, la condizione del corollario 2.9 è anche sufficiente per la convergenza:

Teorema 2.10 (di Reich-Ostrowski). *Se la matrice A è hermitina e definita positiva, condizione necessaria e sufficiente affinchè* $\rho(H_\omega)<1$ *è* $0<\omega<2$.

Dim. La matrice di splitting del metodo SOR è $N=(\frac{D}{\omega} + L)$per cui la matrice Q del lemma 2.6 è:

$$Q = \frac{D}{\omega} + L + \frac{D}{\omega} + L^H - (L + D + L^H) = D\left(\frac{2}{\omega} - 1\right)$$

che risulta definita positiva per $0<\omega<2$.

Metodi iterativi a blocchi.

Quando la matrice del sistema da risolvere presenta una struttura a blocchi, quale quella generata dalla discretizzazione del problema di Dirichlet, i metodi iterativi ora visti possono essere implementati a blocchi. Per fissare le idee, riferiamoci proprio all'esempio citato la cui matrice rappresenteremo ora con la seguente struttura *tridiagonale a blocchi:*

$$A = \begin{pmatrix} D_1 & A_1 & & \\ C_2 & D_2 & A_2 & \\ & C_3 & D_3 & A_3 \end{pmatrix}$$

dove i blocchi diagonali D_1,D_2,D_3,D_4 sono quadrati ed hanno dimensione, rispettivamente, 3,4,3,2, mentre i blocchi sopra e sottodiagonali A_i e C_i sono rettangolari ed inoltre $A_i=C_{i+1}^T$ per i=1,2,3. Ripartendo anche il vettore delle incognite $x=u(P_i)$ e dei

termini noti $f=f(P_i)$ $i=1,...,12$ in 4 sottovettori $x=(x_1,x_2,x_3,x_4)$ ed $f=(f_1,f_2,f_3,f_4)$ di dimensioni uguali alle dimensioni dei blocchi diagonali, il sistema assume la seguente forma "tridiagonale a blocchi"

$$
\begin{aligned}
D_1x_1+A_1x_2 &= f_1\\
C_2x_1+D_2x_2+A_2x_3 &= f_2\\
C_3x_2+D_3x_3+A_3x_3 &= f_3\\
C_4x_3+D_4x_4 &= f_4
\end{aligned}
$$

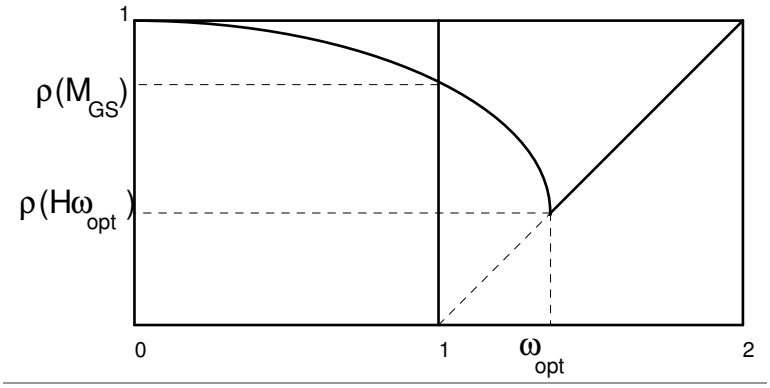
A tale sistema a blocchi, rappresentato formalmente come un sistema usuale, si possono applicare i metodi iterativi visti che si chiameranno metodo di "Jacobi, Gauss-Seidel ed SOR a blocchi". Ciascuno di loro richiede, al blocco i-esimo, la risoluzione di un sistema di matrice D_i.

Per le matrici tridiagonali a blocchi vale il seguente teorema analogo al teorema 2.5 che si riferiva ai metodi iterativi che ora chiameremo *puntuali* per distinguerli da quelli a blocchi.

Teorema 2.11. *Se la matrice* A *è tridiagonale a blocchi, allora per le matrici di iterazione di Jacobi e di Gauss-Seidel a blocchi vale la seguente relazione* $\rho(M_{GS})=\rho^2(M_J)$.

Concludiamo il paragrafo enunciando un teorema relativo alle matrici A hermitiane, definite positive e tridiagonali a blocchi che riassume in parte i teoremi già enunciati e che fornisce esplicitamente il valore del parametro ottimale di rilassamento in funzione del raggio spettrale della matrice di iterazione di Gauss-Seidel.

Teorema 2.12. *Se* A *è una matrice hermitiana, definita positiva e tridiagonale a blocchi, allora, per i metodi puntuali ed a blocchi, si ha* $\rho(M_{GS})=\rho^2(M_J)<1$. *Inoltre i metodi* SOR *puntuali ed* SOR *a blocchi, convergono per ogni* $0<\omega<2$ *ed il valore ottimale del parametro è* $\omega_{opt}=\dfrac{2}{1+\sqrt{1-\rho(M_{GS})}}>1$ *al quale corrisponde un raggio spettrale* $\rho(H_{\omega_{opt}}) = \omega_{opt}-1$. *Il grafico di* $\rho(H_\omega)$ *è dato, qualitativamente, dalla seguente figura*:



Abbiamo già osservato che il sistema relativo al problema di Dirichelet, presentato all'inizio, può essere risolto col metodo di Gauss-Seidel puntuale. Il teorema precedente ci assicura anche la convergenza del metodo di Gauss-Seidel a blocchi e di Jacobi a blocchi. In particolare quello di Jacobi è più lento.

Criteri d'arresto.

Per una effettiva implementazione dei metodi iterativi appena visti, è necessario un criterio d'arresto cioè un meccanismo automatico che interrompa il processo iterativo in base ad una stima dell'errore. Le stime dell'errore relative alla iterazione k-esima possono essere *stime a priori* oppure *stime a posteriori*. Le prime sono fondate solo sui dati del problema e sul punto iniziale dell'iterazione e danno una stima, a priori, del numero di iterazioni necessarie per approssimare la soluzione con un errore inferiore ad una tolleranza assegnata. Le seconde invece fanno uso anche dei valori forniti da ciascuna iterazione e sono quindi, presumibilmente, migliori.

Per una stima a posteriori, si osservi che

$$||x_{k+1}-x|| \le ||M||\ ||x_k-x|| = ||M||\ ||x_k-x_{k+1}+x_{k+1}-x|| < ||M||(||x_k-x_{k+1}||+||x_{k+1}-x||)$$

$$(1-||M||)||x_{k+1}-x|| \le ||M||\ ||x_k-x_{k+1}||$$

$$||x_{k+1}-x|| \le \frac{||M||}{1-||M||}\ ||x_k-x_{k+1}||.$$

Disponendo di una valutazione di ||M|| (<1) e memorizzando ad ogni passo gli ultimi due valori della traiettoria si ottiene, ad ogni iterazione, una stima *a posteriori* dell'errore.

Una stima *a priori* ricavata, in funzione del primo passo della traiettoria x_1-x_0, si può ottenere dalla precedente nel seguente modo.

Si osservi che

$$x_k - x_{k+1} = M(x_{k-1} - x_k) = \ldots = M^k(x_0 - x_1)$$

e che

$$||x_k - x_{k+1}|| \leq ||M||^k ||x_1 - x_0||.$$

Sostituendo quest'ultima nella precedente stima a posteriori si ottine:

$$||x - x_{k+1}|| \leq \frac{1}{1-||M||} ||M||^{k+1} ||x_1 - x_0||.$$

Da una stima di $||M||$ (<1) e dalla conoscienza di x_1 e x_0 si può stimare a priori quante iterazioni occorrono per avere $||x - x_k|| < TOL$.

Un'altro criterio di arresto è fondato sul residuo $r_k := Ax_k - b$ per il quale

$$A^{-1} r_k = x_k - A^{-1} b = x_k - x$$

da cui

$$||x_k - x|| \leq ||A^{-1}|| \, ||r_k||.$$

Per una valutazione corretta dell'errore bisognerebbe disporre della quantità $||A^{-1}||$. In generale si può dire che, siccome $Ax=b$, si ha $||b|| \leq ||A|| \, ||x||$ e $1/||x|| \leq ||A||/||b||$ e quindi

$$\frac{||x_k - x||}{||x||} \leq ||A^{-1}|| \, ||A|| \, \frac{||r_k||}{||b||}.$$

Ciò indica che un piccolo valore relativo del residuo assicura un piccolo valore dell'errore relativo solo per sistemi ben condizionati.

Esempio:

Consideriamo il sistema

$$\begin{cases} x + y = 2 \\ x + 1.01y = 2.01 \end{cases}$$

e supponiamo di aver calcolato una soluzione approssimata

$$\bar{z} = (\bar{x}, \bar{y}) = (10,-8)$$

che fornisce un residuo r (0; -0.09), $||r||_\infty$= 0.09.
Per il termine noto b si ha $||b||_\infty$= 2.01 e quindi $\frac{||r||_\infty}{\|b\|_\infty} \approx 0.045$.

La soluzione vera invece è: z=(1,1). L'errore relativo è quindi $\frac{||z-\bar{z}||_\infty}{\|z\|_\infty} = 9$ che risulta circa 200 volte maggiore di $\frac{||r||_\infty}{\|b\|_\infty}$

2- METODI DIRETTI

I metodi diretti per la risoluzione numerica dei sistemi lineari consistono sostanzialmente nell'applicazione del metodo di riduzione di Gauss con il quale, attraverso la sostituzione di ogni riga con opportune combinazioni lineari della stessa riga con altre, si perviene ad un sistema equivalente di forma triangolare e quindi di immediata risoluzione.

Si badi che, sebbene la soluzione esatta del sistema

$$Ax=b$$

si esprima con

$$x=A^{-1}b,$$

c'è una differanza sostanziale tra il risolvere il sistema, cioè trovare x, e il calcolare la matrice inversa A^{-1}. Basti pensare all'equazione lineare

$$7x=21$$

la cui soluzione è

$$x = \frac{21}{7} = 3$$ (con qualunque precisione di macchina)

e richiede una sola operazione, mentre attraverso il calcolo dell'inversa 1/7 si ottiene la soluzione

$$x = \frac{1}{7}\,21 = 0.142857 \times 21 = 2.99997$$

che richiede due operazioni anzichè una e rivela, di conseguenza, una maggiore propagazione dell'errore.

Il calcolo dell'inversa A^{-1} equivale infatti a risolvere il sistema Ax=b *per tutti i termini noti* b o, più precisamente, data la linearità di R^n, per n termini noti indipendenti. Infatti, poichè l'inversa A^{-1} soddisfa l'equazione matriciale AX=I, le colonne c_i di X sono ottenute risolvendo gli n sistemi

$$Ac_i=e_i \qquad i=1,...,n.$$

Dovendo invece risolvere il sistema Ax=b più di n volte per valori diversi del termine noto b, allora conviene disporre dell'inversa.

Osserveremo comunque che la riduzione a forma triangolare equivale, anche in termini di numero di operazioni, alla fattorizzazione A=LU con L triangolare inferiore ed U triangolare superiore che, una volta ottenuta, può essere utilizzata per la risoluzione di Ax=b per ogni b.

Metodo di riduzione di Gauss e fattorizzazione LU.

Consideriamo il problema Ax=b e definiamo la seguente successione di sistemi equivalenti

$$A^{(i)}x=b^{(i)} \qquad i=1,...,n-1$$

Fissato $A^{(1)}:=A$ e $b^{(1)}:=b$, il sistema di partenza è

$$\begin{aligned}
a_{11}^{(1)}x_1 + \quad a_{12}^{(1)}x_2 + \quad . \ . \ . \quad + a_{1n}^{(1)}x_n \quad &= b_1^{(1)} \\
a_{21}^{(1)}x_1 + \quad a_{22}^{(1)}x_2 + \quad . \ . \ . \quad + a_{2n}^{(1)}x_n \quad &= b_2^{(1)} \\
. \ . \ . & \\
a_{n1}^{(1)}x_1 + \quad a_{n2}^{(1)}x_2 + \quad . \ . \ . \quad + a_{nn}^{(1)}x_n \quad &= b_n^{(1)}
\end{aligned}$$

dal quale si ottine l'equazione $A^{(2)}x=b^{(2)}$ nel seguente modo.
Con un opportuno scambio di righe nel sistema ci si assicura preliminarmente che $a_{11}^{(1)} \neq 0$ e si definiscono quindi i *moltiplicatori*

$$m_{21} = -\frac{a_{21}^{(1)}}{a_{11}^{(1)}} \quad \ldots, \quad m_{n1} = -\frac{a_{n1}^{(1)}}{a_{11}^{(1)}} \quad .$$

Mentre la prima riga del nuovo sistema rimane inalterata, per ogni riga sottostante si esegue la seguente sostituzione che ha lo scopo di annullare i coefficienti $a_{21},a_{31},...,a_{n1}$ della prima colonna di $A^{(2)}$:

$$a_{ij}^{(2)} = a_{ij}^{(1)} + a_{ij}^{(1)}\, m_{i1} \qquad i=2,\ldots,n \qquad j=1,2,...,n$$
$$b_{i}^{(2)} = b_{1}^{(1)}\, m_{i1} + b_{i}^{(1)} \qquad i=2,\ldots,n$$

All'atto pratico le precedenti sostituzioni saranno eseguite solo per j=2,..,n perchè per j=1 già sappiamo che i coefficienti $a_{i1}^{(2)}$ i=2,..,n risulteranno nulli.

Il nuovo sistema $A^{(2)}x=b^{(2)}$ assume quindi la forma

$$\begin{aligned}
&a_{11}^{(1)}x_1 + \quad a_{12}^{(1)}x_2 + \quad \ldots \quad + a_{1n}^{(1)}x_n \quad = b_1^{(1)} \\
&\qquad\quad a_{22}^{(2)}x_2 \quad \ldots \quad + x_n \qquad\quad = b_2^{(2)} \\
&\qquad\quad + \\
&\\
&\qquad\quad a_{n2}^{(2)}x_2 \quad \ldots \quad + a_{nn}^{(2)}x_n \quad = b_n^{(2)} \\
&\qquad\quad +
\end{aligned}$$

Osserviamo che il coefficiente $a_{22}^{(2)}$ ed i coefficienti sottostanti $a_{i2}^{(2)}$ i=3,..,n non possono essere tutti nulli inquanto il determinante di $A^{(2)}$, che coincide con quello di A, è dato dal prodotto di $a_{11}^{(1)}$ per il determinante del minore $A_{11}^{(2)}$ che, di conseguenza, non può essere nullo.

Siamo di nuovo in grado di effettuare uno scambio delle righe sul sistema $A^{(2)}x=b^{(2)}$ che porti nella posizione di indice (2,2) (che d'ora in poi chiameremo *posizione di pivot* della matrice $A^{(2)}$) un coefficiente non nullo.

Ottenuto il sistema $A^{(k)}x=b^{(k)}$ del tipo

$$
\begin{array}{lllll}
a_{11}^{(1)} x_1 + \ldots & a_{1k}^{(1)} x_k + \ldots\ldots & + a_{1n}^{(1)} x_n & = b_1^{(1)} \\
\vdots & & \vdots & \vdots \\
 & a_{kk}^{(k)} x_k + a_{k;k+1}^{(k)} x_{k+1} + \ldots & + a_{kn}^{(k)} x_n & = b_k^{(k)} \\
 & a_{k+1;k+1}^{(k)} x_{k+1} + \ldots & + a_{k+1;n}^{(k)} x_n & = b_{k+1}^{(k)} \\
 & \vdots & \vdots & \\
 & a_{n;k+1}^{(k)} x_{k+1} + \ldots & + a_{nn}^{(k)} x_n & = b_n^{(k)}
\end{array}
$$

si porta in posizione di pivot (k+1,k+1) un coefficiente non nullo (attraverso un eventuale scambio della riga (k+1)-esima con una riga sottostante), poi si calcolano i moltiplicatori

$$
m_{i,k+1} = - \frac{a_{i;k+1}^{(k)}}{a_{k+1;k+1}^{(k)}} \qquad i=k+2,\ldots,n \qquad (*)
$$

e, successivamente, si effettuano le sostituzioni

$$
\begin{aligned}
a_{i\,j}^{(k+1)} &= a_{i\,j}^{(k)} + a_{k+1;j}^{(k)}\, m_{i,k+1} \qquad i=k+2,\ldots,n, \qquad j=k+1,\ldots,n \qquad (*) \\
b_i^{(k+1)} &= b_{k+1}^{(k)}\, m_{i,k+1} + b_i^{(k)} \qquad i=k+2,\ldots,n,
\end{aligned}
$$

Al passo (n-1)-esimo si trova infine il sistema $A^{(n)}x=b^{(n)}$

$$
\begin{array}{lll}
a_{11}^{(1)} x_1 + \ldots + & a_{1n}^{(1)} x_n & = b_i^{(1)} \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
 & a_{nn}^{(n)} x_n & = b_n^{(n)}
\end{array}
$$

la cui matrice $A^{(n)}$ è di tipo triangolare superiore e indicheremo con U.
A questo punto è immediato calcolare le componenti di x all'indietro da x_n ad x_1.

Osserviamo che, a parte eventuali scambi di righe, la matrice $A^{(2)}$ è stata ottenuta da $A^{(1)}$ attraverso la moltiplicazione

$$
A^{(2)}=M_1 A^{(1)}
$$

dove

$$M_1 = \begin{Vmatrix} 1 & & & \\ m_{21} & 1 & & \\ . & & . & \\ . & & & . \\ m_{n1} & . \ . \ . & & 1 \end{Vmatrix}$$

e successivamente,

$$A^{(3)}=M_2A^{(2)}=M_2M_1A^{(1)}$$

.....

$$A^{(n)}=M_{n-1}A^{(n-1)}=M_{n-1}...M_1A^{(1)}=U$$

dove, in generale,

$$M_k = \begin{Vmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & m_{k+1,k} & 1 & \\ & & : & & . \ . \\ & & m_{n,k} & & 1 \end{Vmatrix}$$

Siccome le matrici M_k sono tutte invertibili, abbiamo fattorizzato la matrice A in $A=(M_{n-1}...M_1)^{-1}U=M_1^{-1}...M_{n-1}^{-1}U$.

Le matrici del tipo M_k, dette **matrici elementari di Gauss**, sono esprimibili come somma dell'identità e di una matrice di rango 1:

$$M_k=I+m_ke_k^T \quad \text{dove} \quad m_k=(0,....0,m_{k+1,k},...,m_{n,k})^T$$

Si osservi che la matrice $m_k e_k^T$ è nihilpotente $\left((m_k e_k^T)^2=0\right)$ cosicchè la serie di Neumann è $(I+m_k e_k^T)^{-1}=I-m_k e_k^T$. Si ha quindi la seguente espressione per l'inversa M_k^{-1}

$$M_k^{-1} = \begin{Vmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & -m_{k+1,k} & 1 & & \\ & & : & & \ddots & \\ & & -m_{n,k} & & & 1 \end{Vmatrix}$$

e per il prodotto $M_k^{-1} M_{k+1}^{-1} = (I-m_k e_k^T)(I-m_k e_{k+1}^T) = I - m_k e_k^T - m_k e_{k+1}^T$

$$M_k^{-1} \quad M_{k+1}^{-1} \quad = \begin{Vmatrix} 1 & \ddots & & & & \\ & \ddots & 1 & & & \\ & & -m_{k+1;k} & 1 & & \\ & & & -m_{k+2,k+1} & & \\ & & : & : & \ddots & \\ & & -m_{n,k} & -m_{k+2,k+1} & & 1 \end{Vmatrix}$$

Cosicchè è facile vedere che la matrice $L_{k+1}:=M_1^{-1}...M_{k+1}^{-1}$ è

$$L_{k+1} \quad = \begin{Vmatrix} 1 & & & & & \\ -m_{21} & 1 & & & & \\ : & & 1 & & & \\ & & \ddots & -m_{k+2,k+1} & \ddots & \\ : & & : & & \ddots & \\ -m_{n1} & \ddots & \ddots & -m_{k+2,k+1} & & 1 \end{Vmatrix}$$

ed è ottenuta da L_k semplicemente riempiendo la parte sottodiagonale della colonna (k+1)-esima con i nuovi moltiplicatori cambiati di segno.

Infine la matrice $L:=L_{n-1}=M_1^{-1}...M_{n-1}^{-1}$ è ancora di tipo triangolare inferiore con diagonale unitaria:

$$L_{n-1} = \begin{Vmatrix} 1 & & & \\ -m_{21} & 1 & . & \\ . & & . & \\ . & & & 1 \\ -m_{n1} & \dots & -m_{n,n-1} & 1 \end{Vmatrix}$$

Abbiamo dunque ottenuto la fattorizzazione

$$A = L\,U$$

con L ed U triangolari inferiore e superiore e con $\det(A)=\det(U)$.

Analizziamo ora gli scambi di righe che abbiamo ipotizzato di fare nei precedenti passaggi. Analizziamo, per semplicita', quello che accade quando si fa uno scambio solo. Supponiamo dunque che, una volta ottenuto il sistema $A^{(k)}x=b^{(k)}$, prima di calcolare $A^{(k+1)}$ si esegua lo scambio della riga (k+1)-esima con la i-esima $(i>k+1)$ attraverso la matrice di permutazione $P_{i,k+1}$:

$$A^{(k)} \longrightarrow P_{i,k+1}A^{(k)} = P_{i,k+1}M_k...M_1A$$

cosicchè

$$A^{(k+1)}=M_{k+1}P_{i,k+1}M_k...M_1A$$

e quindi

$$A=M_1^{-1}...M_k^{-1}\,P_{i,k+1}M_{k+1}^{-1}\,A^{(k+1)} = L_kP_{i,k+1}M_{k+1}^{-1}\,A^{(k+1)}$$

Applicando la permutazione $P_{i,k+1}$ ad entrambi i membri si trova

$$P_{i,k+1}A=P_{i,k+1}L_k\,P_{i,k+1}M_{k+1}^{-1}\,A^{(k+1)}$$

dove la matrice

$$P_{i,k+1}\,L_k\,P_{i,k+1}$$

differisce da L_k solo per lo scambio dei moltiplicatori delle righe (k+1)-esima ed i-esima. Ciò significa che se la matrice $A^{(k)}$ necessita di una certa permutazione di righe $P_{i,k+1}$, la stessa permutazione va effettuata sulle corrispondenti righe della matrice dei moltiplicatori L_k e la fattorizzazione LU che si ottiene alla fine del processo di triangolarizzazione è relativa ad una permutazione P di A che tenga conto di tutte le permutazioni effettuate.

$$L\,U = P\,A$$

In conclusione il sistema da risolvere sarà

$$LUx=Pb$$

la cui soluzione è direttamente ottenibile dalla risoluzione dei due sistemi triangolari

$$Ly=Pb \qquad e \qquad Ux=y.$$

Analizzando le formule (*), osserviamo ancora che l'ordine con cui vengono costruiti i coefficienti di L e di U è il seguente: (**processo di "pavimentazione"**)

1ª riga di U ; 1ª colonna di L ; 2ª riga di U ; 2ª colonna di L ; ... nª riga di U.

Pertanto le colonne di L e le righe di U possono essere allocate, man mano che si calcolano, al posto delle corrispondenti colonne e righe di A i cui coefficienti non sono più necessari (essendo la diagonale di L unitaria, non occorre memorizzarla). Ciò significa che la fattorizzazione LU non richiede ulteriore occupazione di memoria oltre quella necessaria per la memorizzazione di A.

Concludiamo questo paragrafo con il seguente teorema di unicità.

TEOREMA 2.9. *Ogni matrice* A *non singolare è fattorizzabile in modo unico con due matrici* L *ed* U *tali che* A=LU *dove* L *è triangolare inferiore a diagonale unitaria ed* U *triangolare superiore*.

Dim. Se fosse A=LU ed A=L'U', sarebbe:

$LU=L'U'$
$U=L^{-1}L'U'$
$UU'^{-1}=L^{-1}L'$.

Si osservi ora che $L^{-1}L'$ è ancora triangolare inferiore con diagonale unitara, mentre UU'^{-1} è triangolare superiore. Essendo tra loro uguali devono coincidere con l'identità, di conseguenza L=L' ed U=U'.

Strategia del pivot.

Abbiamo visto che al passo k-esimo della fattorizzazione il pivot $a_{k,k}^{(k-1)}$ deve essere $\neq 0$. La strategia del **pivot parziale** consiste nell'eseguire in ogni caso uno scambio di righe che porti nella posizione di pivot il coefficiente di modulo massimo della colonna sottostante. La strategia del **pivot totale** consiste invece nell'eventuale scambio anche delle colonne successive in modo da portare nella posizione di pivot il coefficiente di modulo massimo tra tutti quelli del minore definito dagli elementi sottostanti: $a_{i,j}^{(k-1)}$ con $i,j \geq k$.

A differenza del pivot parziale, la strategia del pivot totale richiede, ad ogni scambio di colonne, il riordinamento delle incognite. La strategia del pivot oltre ad evitare il rischio di una divisione per zero, consente di ridurre, come vedremo più avanti, la propagazione dell'errore dovuta al gran numero di operazioni richieste per la fattorizzazione.

In presenza di sistemi ben condizionati, la strategia del pivot non sarebbe strettamente necessaria se non per escludere, come già osservato, il rischio di una divisione per zero. Allora ci si chiede se è possibile stabilire a priori che ad ogni passo del processo di fattorizzazione il pivot sia non nullo. A questo proposito valgono i seguenti teoremi:

TEOREMA 2.10. *Se i minori principali di* A *sono non singolari allora, per ogni matrice* $A^{(k)}$, *è* $a^{(k)}_{k+1;k+1} \neq 0$.

Dim. L'asserto è ovviamente vero per k=0 e supponiamolo vero per k. Poichè i minori principali di A e di $A^{(k)}$ hanno lo stesso determinante, anche il minore principale (k+1)-esimo di $A^{(k)}$ è non singolare e quindi $a^{(k)}_{k+1;k+1} \neq 0$.

$$\begin{Vmatrix} a^{(k)}_{11} & \cdots & a^{(k)}_{1k} & \cdots & \cdots & a^{(k)}_{1n} \\ 0 & \ddots & & & & \vdots \\ & & a^{(k)}_{kk} & a^{(k)}_{k;k+1} & \cdots & a^{(k)}_{kn} \\ 0 & \cdots & 0 & a^{(k)}_{k+1;k+1} & \cdots & a^{(k)}_{k+1;n} \\ & & & \vdots & & \vdots \\ 0 & & 0 & a^{(k)}_{n;k+1} & \cdots & a^{(k)}_{nn} \end{Vmatrix}$$

TEOREMA 2.11 . *Se* A *ha predominanza diagonale stretta oppure è definita positiva, allora i minori principali sono non singolari*.

Dim. Sia A a predominanza diagonale, allora lo è ogni minore principale che, per il teorema di Gerschgorin, risulta pertanto non singolare. Sia invece A definita positiva; allora lo è ogni minore principale. Sia infatti A_i il minore principale i-esimo e sia $y_i := (x_1,...,x_i) \in R^i$.
Detto inoltre $u=(x_1,...,x_i,0,...,0) \in R^n$ si ha

$$y^{T,i} A_i\, y = u^T A\, u > 0$$

e quindi $\det(A_i) \neq 0$.

Complessità computazionale.

Il confronto di algoritmi diversi per la risoluzione di uno stesso problema è, in generale, un compito assai arduo, ancorchè possibile, in quanto la prestazione di un algoritmo dipende in larga misura dal modo in cui esso viene implementato, dal linguaggio di programmazione usato e dalla macchina sulla quale viene eseguito. Non è negli obbiettivi di questo corso fare una analisi così raffinata degli algoritmi e quindi ci accontenteremo di enumerare semplicemente il numero di operazioni necessarie per realizzare l'algoritmo in oggetto, valuteremo cioè la sua **complessità computazionale**. In

ogni caso assumeremo come operazione elementare il complesso di una moltiplicazione (o divisione) ed una successiva addizione sul risultato.

Il passaggio dalla matrice $A^{(1)}$ alla matrice $A^{(2)}$ richiede, per ogni riga dalla 2^a alla n^a, il calcolo del moltiplicatore m_{i1} (1 op) e delle quantità

$$a_{ij}^{(2)} = a_{ij}^{(1)} + a_{1j}^{(1)}\, m_{i1} \qquad j=2,...,n \qquad (n-1 \text{ op})$$

per un totale di n op. La matrice L_1 ed $A^{(2)}$ richiedono quindi $n(n-1)$ op. Analogamente per il passaggio da $A^{(2)}$ ad $A^{(3)}$ bisogna manipolare la sottomatrice $B^{(1)}$ di dimensione $n-1$:

$$A^{(2)} = \begin{Vmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & & \\ & B^{(1)} & \\ 0 & & \end{Vmatrix}$$

per un totale di $(n-1)(n-2)$ op. Infine il passaggio da $A^{(n-2)}$ ad $A^{(n-1)}$ richiede 2 op. L'intera trasformazione, cioè la determinazione di L ed U, richiede

$$\sum_{j=1}^{n-1} j(j+1) = \sum_{j=1}^{n-1} j^2 + \sum_{j=1}^{n-1} j = \frac{(n-1)n(2n-1)}{6} + \frac{n(n-1)}{2}$$

$$= \frac{n^3}{3} - \frac{n}{3} \approx \frac{n^3}{3}$$

Infine per calcolare x devo risolvere i due sistemi Ly=b e Ux=y .

Per Ly=b:

$y_1 = b_1$	0 op
$y_2 = b_2 - m_{21}y_1$	1 op
$y_3 = b_3 - m_{31}y_1 + m_{32}y_2$	2 op
...	
$y_n = b_n - m_{n1}y_1 - \ldots - m_{n,n-1}y_{n-1}$.	n-1 op

in totale $n(n-1)/2$ op.

Per Ux=y:

$x_n = y_n/u_{nn}$	1 op
$x_{n-1} = (y_{n-1} - u_{n-1,n}x_n)/u_{n-1,n-1}$	2 op
...	

$x_1=(y_1 - u_{1,n}x_n -... -u_{12}x_2)/u_{11}$ n op

in totale n(n+1)/2 op.

I due sistemi triangolari richiedono globalmente n^2 op. Asintoticamente il tempo richiesto per la loro risoluzione è trascurabile rispetto al tempo necessario per la fattorizzazione LU.

Metodo di Choleski.

Nel caso in cui la matrice A sia simmetrica e definita positiva la fattorizzazione con matrici triangolari inferiore e superiore può essere fatta in modo più economico. Vale infatti il seguente teorema:

Teorema 2.12. *Sia* A *hermitiana e definita positiva, allora esiste una ed una sola matrice triangolare inferiore* $\bar{L}$ *tale che* $A=\bar{L}\,\bar{L}^H$.

Dim. Sia LU la fattorizzazione di A e sia $D:=diag(u_{11}....u_{nn})$. Si ha quindi:

$A=LDD^{-1}U = LDU'$ con $U'=D^{-1}U$ triangolare superiore a diagonale unitaria

$A=A^H = (U')^H DL^H$

e, per l'unicità della fattorizzazione (con L a diagonale unitaria) si ha:

$(U')^H = L$

e quindi:

$A= LDL^H$.

Detto infine

$D^{1/2}:=diag(\sqrt{u_{11}},...,\sqrt{u_{nn}})$ e $\bar{L}:=LD^{1/2}$,

si ha

$$A = LD^{1/2}D^{1/2}L^H = \bar{L}\,\bar{L}^H.$$

Sul piano pratico, la fattorizzazione $\bar{L}\,\bar{L}^H$ si realizza direttamente uguagliando riga per riga il prodotto $\bar{L}\,\bar{L}^H$ con la matrice A. Si ottine così, per la prima riga di A (supponiamo, per semplicità di notazione, che $\bar{L}$ sia reale):

$$\ell_{11}(\ell_{11},\ell_{21},...,\ell_{2n}) = (a_{11},a_{12},...,a_{1n})$$

dalla quale si ricava la prima riga di $\bar{L}^H$ (cioe' la prima colonna di L):

$\ell_{11}=\sqrt{a_{11}}$
$\ell_{21}=a_{12}/\ell_{11}$
...
$\ell_{2n}=a_{1n}/\ell_{11}$.

Analogamente per la seconda riga di A :

$$\ell_{21}(\ell_{11},\ell_{21},...,\ell_{2n})+\ell_{22}(0,\ell_{22},\ell_{32},...,\ell_{3n})=(a_{21},a_{22},...,a_{2n})$$

dalla quale si ricava la seconda riga di L^H:

$\ell_{22}=\sqrt{a_{22} - \ell_{21}^2}$
$\ell_{32}=a_{23}-\ell_{21}\ell_{31}/\ell_{22}$
...
$\ell_{n2}=a_{2n}-\ell_{21}\ell_{n1}/\ell_{22}$.

E così di seguito per le righe successive.

Analisi dell'errore.

L'analisi dell'errore nel metodo di Gauss è basata essenzialmente sull'analisi all'indietro ideata da Wilkinson per questo problema intorno agli anni 60. Si dimostra che la fattorizzazione $\bar{L}\bar{U}$ computata in una aritmetica di precisione eps soddisfa la relazione

$$\bar{L}\bar{U} = A + E \qquad \text{con } ||E||_\infty \leq n^2 g_n ||A||_\infty \text{ eps}$$

dove

$$g_n = \frac{\max_{i\,j\,k}|\bar{a}_{i\,j}^{(k)}|}{\max_{i\,j}|a_{i\,j}|}$$

Inoltre, detta $\bar{y}$ la soluzione computata di $\bar{L}y=b$ ed $\bar{x}$ quella di $\bar{U}x=\bar{y}$, si dimostra che $\bar{x}$ soddisfa l'equazione:

$$(A + \delta A)\bar{x} = b \qquad \text{con } ||\delta A|| \leq (n^3+3n^2)\, g_n ||A||_\infty \text{ eps.}$$

Obbiettivo di un buon algoritmo è quello di ottenere una fattorizzazione con una perturbazione E più piccola possibile. Ciò dipende essenzialmente dalla costante g_n per la quale si possono dare le seguenti stime:

strategia del pivot parziale: $g_n \leq 2^{n-1}$

strategia del pivot totale: $g_n \leq 1.8\, n^{0.25 \log n}$.

per matrici Hermitiane definite positive: $g_n \leq 1$

Vediamo ora in generale che cosa comporta, per la soluzione, una perturbazione sui dati del sistema, cioè sulla matrice e sul termine noto. Cominciamo col chiederci quando una pertubazione E sulla matrice non singolare A conserva la nonsingolarità per A+E. A questo proposito vale il seguente risultato.

Lemma di perturbazione (di Banach). *Sia* A *non singolare ed* E *tale che* $||A^{-1}||||E||<1$. *Allora* A+E *è ancora non singolare ed inoltre:*

$$||(A+E)^{-1}|| \leq \frac{||A^{-1}||}{1-||A^{-1}||\,||E||}$$

Detta x la soluzione del sistema Ax=b, sia y la soluzione del sistema perturbato

$$(A + \delta A)y = b + \delta b$$

per il quale supponiamo $||A^{-1}||\,||\delta A||<1$. Si avrà quindi:

$$(A + \delta A)x = b + \delta A x$$

$$(A + \delta A)(x - y) = b + \delta A x - b - \delta b = \delta A x - \delta b$$

$$x - y = (A + \delta A)^{-1}(\delta A x - \delta b)$$

$$||x - y|| \le ||(A+\delta A)^{-1}|| \left(||\delta A||\ ||x|| + ||\delta b|| \right) \le \frac{||A^{-1}||}{1 - ||A^{-1}||\ ||\delta A||} \left(||\delta A||\ ||x|| + ||\delta b|| \right).$$

Inoltre:

$$Ax = b \ \Rightarrow ||b|| \le ||A||\ ||x|| \ \Rightarrow \frac{1}{||x||} \le \frac{||A||}{||b||}$$

e quindi:

$$\frac{||x-y||}{||x||} \le \frac{||A^{-1}||}{1 - ||A^{-1}||\ ||\delta A||} \left(||\delta A|| + \frac{||\delta b||\ ||A||}{||b||} \right)$$

$$\frac{||x-y||}{||x||} \le \frac{||A||\ ||A^{-1}||}{1 - ||A^{-1}||\ ||\delta A||} \left(\frac{||\delta A||}{||A||} + \frac{||\delta b||}{||b||} \right)$$

La stima precedente mostra che l'errore relativo causato dalle perturbazioni è maggiorato da una quantità proporzionale alle perturbazioni relative sui dati, con costante di proporzionalità $\frac{||A||\ ||A^{-1}||}{1 - ||A^{-1}||\ ||\delta A||}$ che, in prima approssimazione, vale $||A||\ ||A^{-1}||$. In particolare, se $\delta A = 0$, cioè se la perturbazione riguarda solo il termine noto b, allora si ha:

$$\frac{||x-y||}{||x||} \le ||A||\ ||A^{-1}||\ \frac{||\delta b||}{||b||}$$

Il numero

$$K(A) := ||A||\ ||A^{-1}||$$

è detto **indice di condizionamento** della matrice A e rappresenta dunque la *sensitività* della soluzione di Ax=b rispetto alle perturbazioni sui dati.

L'indice di condizionamento dipende dalla norma e, nel caso estremo di matrice singolare, lo definiamo uguale a infinito:

$$K_i(A) = \begin{cases} ||A^{-1}||_i \; ||A||_i & |A| \neq 0 \\ \infty & |A| = 0 \end{cases}$$

Per ogni norma naturale si ha $K(A) \geq 1$, infatti:

$$1 = ||\mathbb{I}|| = ||A^{-1}A|| \leq ||A^{-1}|| \; ||A||$$

In generale non si riesce a calcolare il valore dell'indice di condizionamento senza disporre dell'inversa A^{-1}, si riesce però a dare la seguente stima:

$$||A^{-1}|| \; ||A|| \geq \rho(A^{-1})\rho(A) = \frac{|\lambda|_{max}}{|\lambda|_{min}}$$

che nel caso di matrici hermitiane diventa, in norma 2,:

$$||A^{-1}||_2 ||A||_2 = \frac{|\lambda|_{max}}{|\lambda|_{min}}.$$

La situazione di condizionamento ottimale si ha nel caso di $K(A)=1$ nel quale la matrice è detta **perfettamente condizionata**. Ciò si verifica, sempre nella norma 2, per le matrici unitarie.

Si osservi infine che nel passaggio dalla matrice A alla matrice $A^H A$ si ha un peggioramento del condizionamento. Infatti , mentre abbiamo visto che per la matrice hermitiana $A^H A$ si ha

$$K_2(A^H A) = \frac{|\lambda_{A^H A}|_{max}}{|\lambda_{A^H A}|_{min}},$$

per la matrice A si ha :

$$||A||_2 = \sqrt{\rho(A^H A)} = \sqrt{|\lambda_{A^H A}|_{max}}$$

e

$$||A^{-1}||_2 = \sqrt{\rho((A^H A)^{-1})} = \sqrt{\frac{1}{|\lambda_{A^H A}|_{min}}}$$

e quindi

$$K_2(A) = \sqrt{\frac{|\lambda_{A^H A}|_{max}}{|\lambda_{A^H A}|_{min}}}.$$

In definitiva gli indici di condizionamento delle due matrici sono legati della relazione $K_2(A^TA)=\left(K_2(A)\right)^2$ e risultano uguali solo per matrici prefettamente condizionate.

Esempio:

A titolo di esempio si consideri il sistema lineare

$$x_1 + x_2 = 2$$
$$x_1 + 1.01x_2 = 2.01$$

la cui soluzione è $x_1=1$, $x_2=1$. La matrice è simmetrica con autovalori $\lambda_1 \cong 2.005$ e $\lambda_2 \cong .005$ per cui $K_2(A) \cong 401$ che rivela un cattivo condizionamento. Infatti il seguente sistema, ottenuto dal precedente con una perturbazione relativa non superiore a .01 sui coefficienti:

$$y_1 + y_2 = 2$$
$$1.001y_1 + y_2 = 2.01$$

ha come soluzione $y_1=10$ e $y_2=-8$, con uno scarto relativo $\frac{||x - y||_2}{||x||_2} = 9$ rispetto ad una perturbazione sulla matrice $\frac{||\delta A||_2}{||A||_2} \approx 0.005$, che risulta amplificata di un fattore 1800. Ciò non è in contraddizione con le maggiorazioni precedenti inquanto esse valgono sotto la condizione

$||\delta A||_2||A^{-1}||_2<1$ che in questo caso non è verificata.

Come già accennato in precedenza (nel paragrafo sui criteri d'arresto per i metodi iterativi) anche il test sul residuo per valutare la bontà di una soluzione approssimata non è accettabile nel caso di sistemi mal condizionati. Proviamo infatti a calcolare il residuo r del primo sistema rispetto ad una soluzione perturbata $y_1=10$ e $y_2=-8$.

$$r_1=y_1 + y_2 - 2 = 10 -8 -2 =0$$
$$r_2=y_1 + 1.01y_2 -2.01=10 -8.08 -2.01= -0.09$$

Il residuo risulta molto "piccolo" rispetto allo scarto relativo sulla soluzione che abbiamo visto essere 9. Ciò è completamente giustificato dall'analisi precedente quando si osservi che il residuo non è altro che una perturbazione sul termine noto. Infatti

$$Ax=b$$

$$Ay=Ay-b+b=r+b$$

Nel nostro caso si ha $\frac{||r||_2}{||b||_2} = \frac{0.045}{\sqrt{2}} \cong 0.0318$ a cui corrisponde un errore relativo $\frac{||x-y||_2}{||x||_2} = 9$ che risulta amplificato di un fattore $\cong 283$.

Raffinamento iterativo

I metodi che abbiamo analizzato in questo paragrafo consistono in un numero finito di operazioni alla fine delle quali si ottiene un risultato che differisce dalla soluzione esatta per il solo effetto degli errori di arrotondamento che si sono propagati durante l'esecuzione dell'intero l'algoritmo. In particolare nel caso della fattorizzazione LU, si ottiene un risultato, che indicheremo ora con x^0, che è la soluzione approssimata del problema

$$\bar{L}\bar{U}x=b$$

con $\bar{L}$ ed $\bar{U}$ a loro volta approssimazioni delle matrici L ed U.

Se la matrice A non è troppo "mal condizionata" si può eseguire il seguente *raffinamento iterativo* della soluzione che consiste nella esecuzione di alcuni passi della iterazione

$$Nx^{i+1}=Px^{i} + b$$

definita dallo splitting $N=\bar{L}\bar{U}$ e $P=\bar{L}\bar{U}-A$, a partire dal valore x^0.

Si ha dunque:

$$\bar{L}\bar{U}x^{i+1}=(\bar{L}\bar{U} -A)x^{i} + b$$

$$\bar{L}\bar{U}(x^{i+1}-x^{i}) = -Ax^{i} + b = r^{i}.$$

Le *correzioni* $x^{i+1}-x^i$ si calcolano, come abbiamo già osservato, attraverso la risoluzione in avanti e all'indietro dei sistemi

$$\bar{L}y = r^i$$

$$\bar{U}(x^{i+1}-x^i) = y$$

nei quali i residui r^i devono essere calcolati in doppia precisione. Ciò è suggerito dal fatto che nella risoluzione del sistema $\bar{L}\bar{U}(x^{i+1}-x^i)=r^i$,una perturbazione relativa sul residuo dell'ordine della precisione di macchina eps, si riflette (se la matrice non è troppo "mal condizionata") in un errore relativo sulla soluzione dello stesso ordine di grandezza. Ora la soluzione x^0 è già stata calcolata con la precisione eps e quindi per avere un miglioramento il residuo deve essere calcolato con una precisione superiore. In certi casi il meccanismo di raffinamento è molto efficace ed un paio di iterazioni sono sufficienti per ottenere una sorprendente precisione.